How is ChatGPT's behavior changing over time?

Year: 2023 | Citations: 529 | Authors: Lingjiao Chen, Matei Zaharia, James Y. Zou

---

Abstract

GPT-3.5 and GPT-4 are the two most widely used large language model (LLM) services. However, when and how these models are updated over time is opaque. Here, we evaluate the March 2023 and June 2023 versions of GPT-3.5 and GPT-4 on several diverse tasks: 1) math problems, 2) sensitive/dangerous questions, 3) opinion surveys, 4) multi-hop knowledge-intensive questions, 5) generating code, 6) US Medical License tests, and 7) visual reasoning. We find that the performance and behavior of both GPT-3.5 and GPT-4 can vary greatly over time. For example, GPT-4 (March 2023) was reasonable at identifying prime vs. composite numbers (84% accuracy) but GPT-4 (June 2023) was poor on these same questions (51% accuracy). This is partly explained by a drop in GPT-4's amenity to follow chain-of-thought prompting. Interestingly, GPT-3.5 was much better in June than in March in this task. GPT-4 became less willing to answer sensitive questions and opinion survey questions in June than in March. GPT-4 performed better at multi-hop questions in June than in March, while GPT-3.5's performance dropped on this task. Both GPT-4 and GPT-3.5 had more formatting mistakes in code generation in June than in March. We provide evidence that GPT-4's ability to follow user instructions has decreased over time, which is one common factor behind the many behavior drifts. Overall, our findings show that the behavior of the"same"LLM service can change substantially in a relatively short amount of time, highlighting the need for continuous monitoring of LLMs.